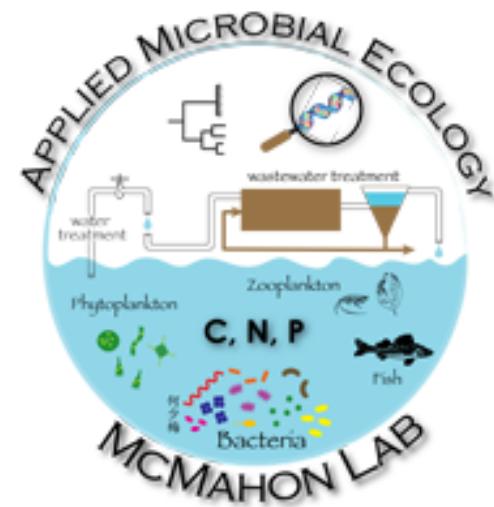
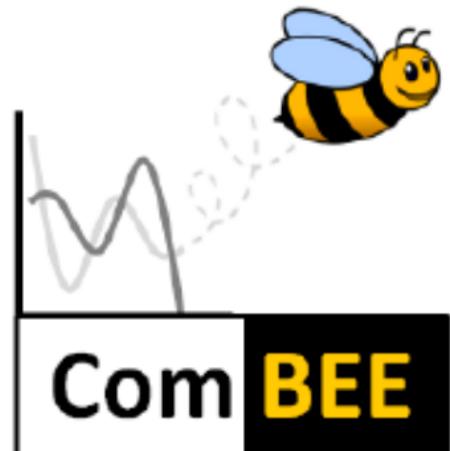


ComBEE 'Omics Study Group: Phylogenetic Analysis

Benjamin Peterson



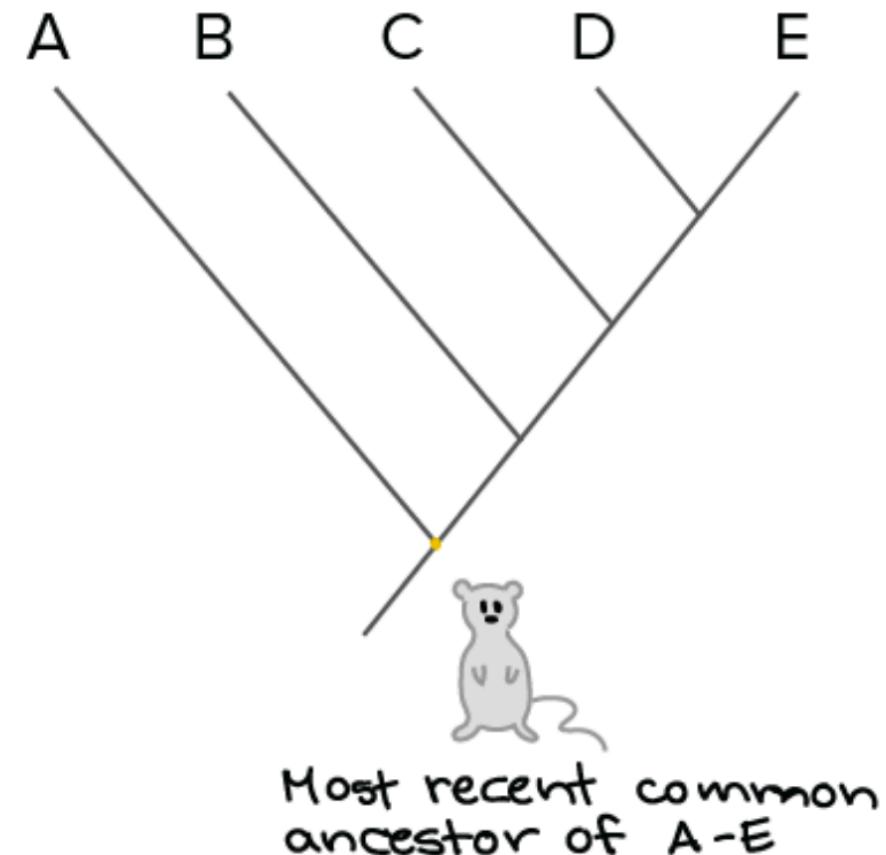
What is phylogeny?

Phûlon – “genus”, “species”, “tribe”

-gèneia – “generation”

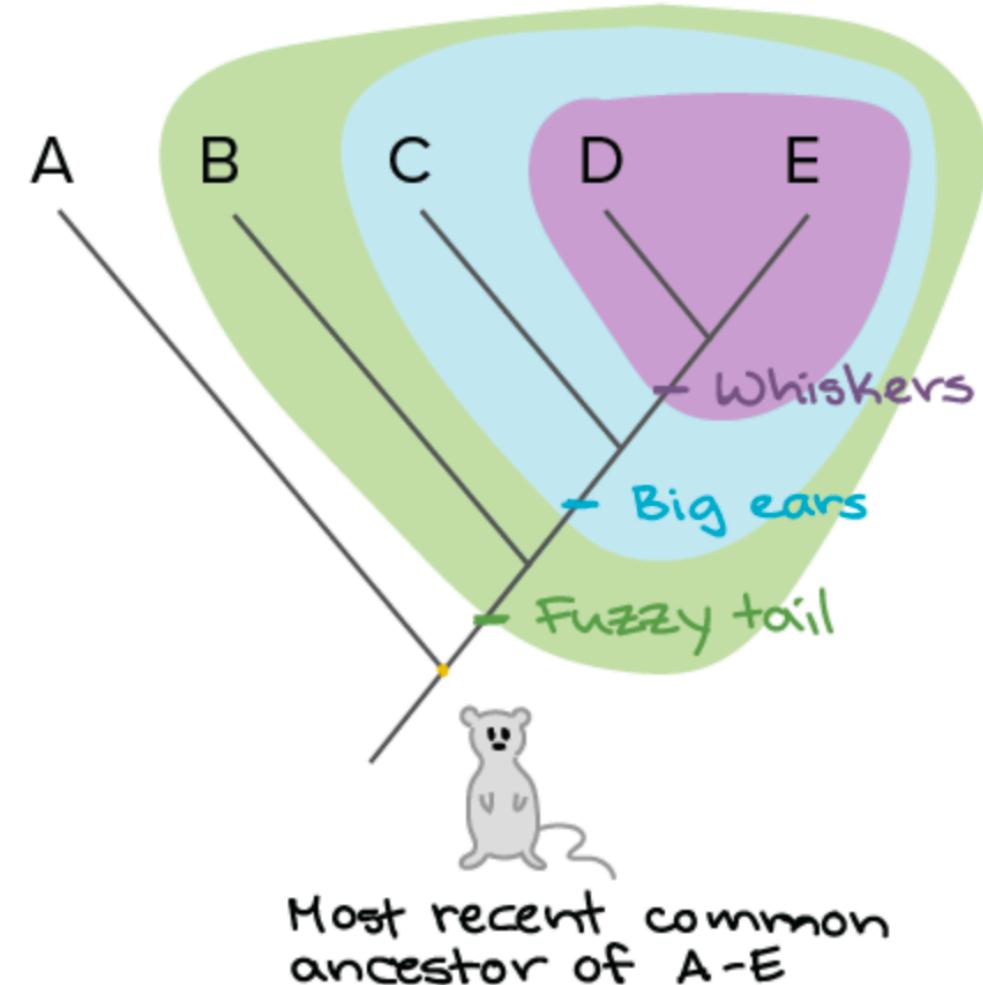
What is phylogenetic inference?

- A hypothesis regarding the evolutionary relationship between units of life
- What is this based on?



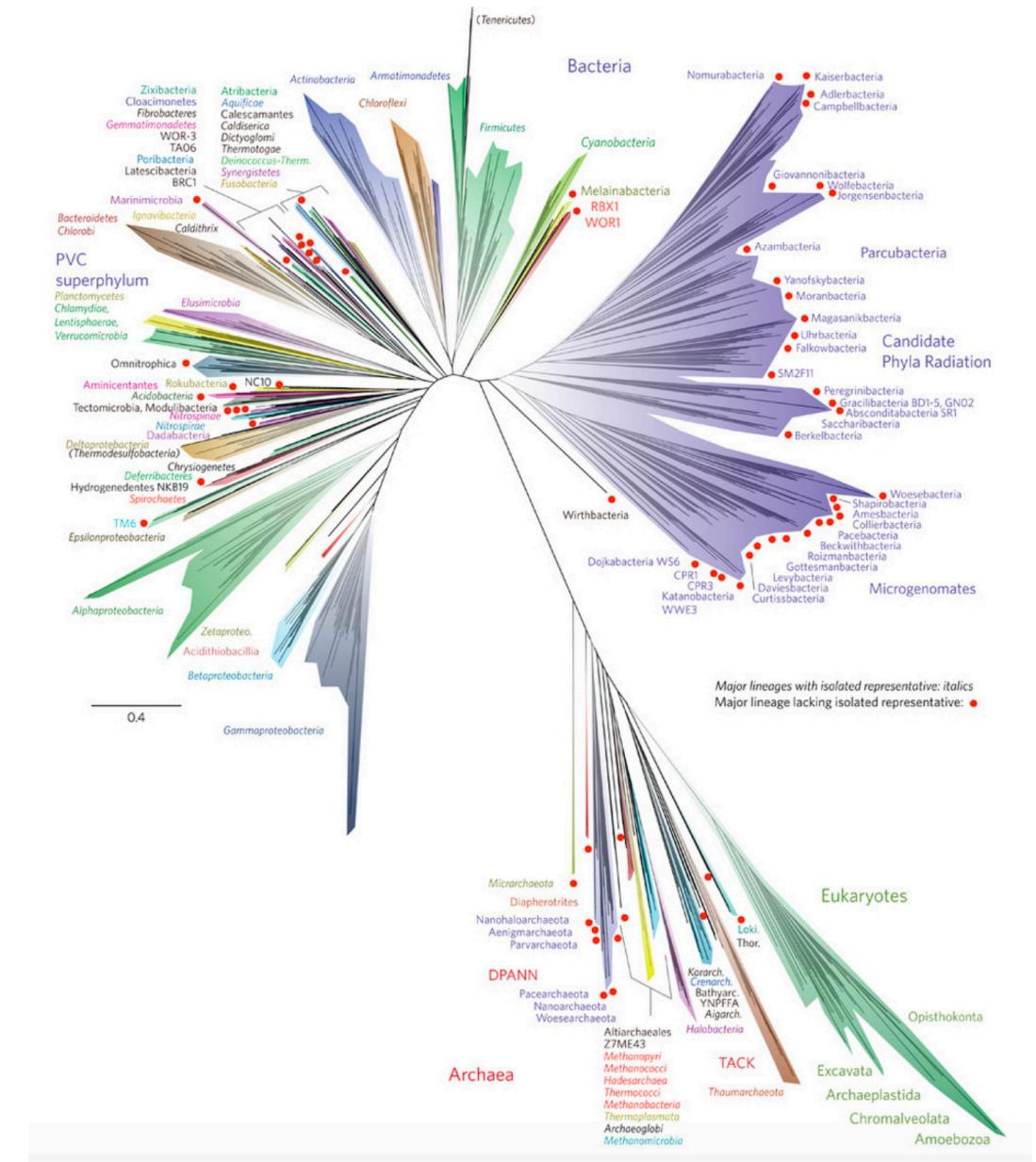
What is phylogenetic inference?

- A hypothesis regarding the evolutionary relationship between units of life
- Primarily based on “derived traits”



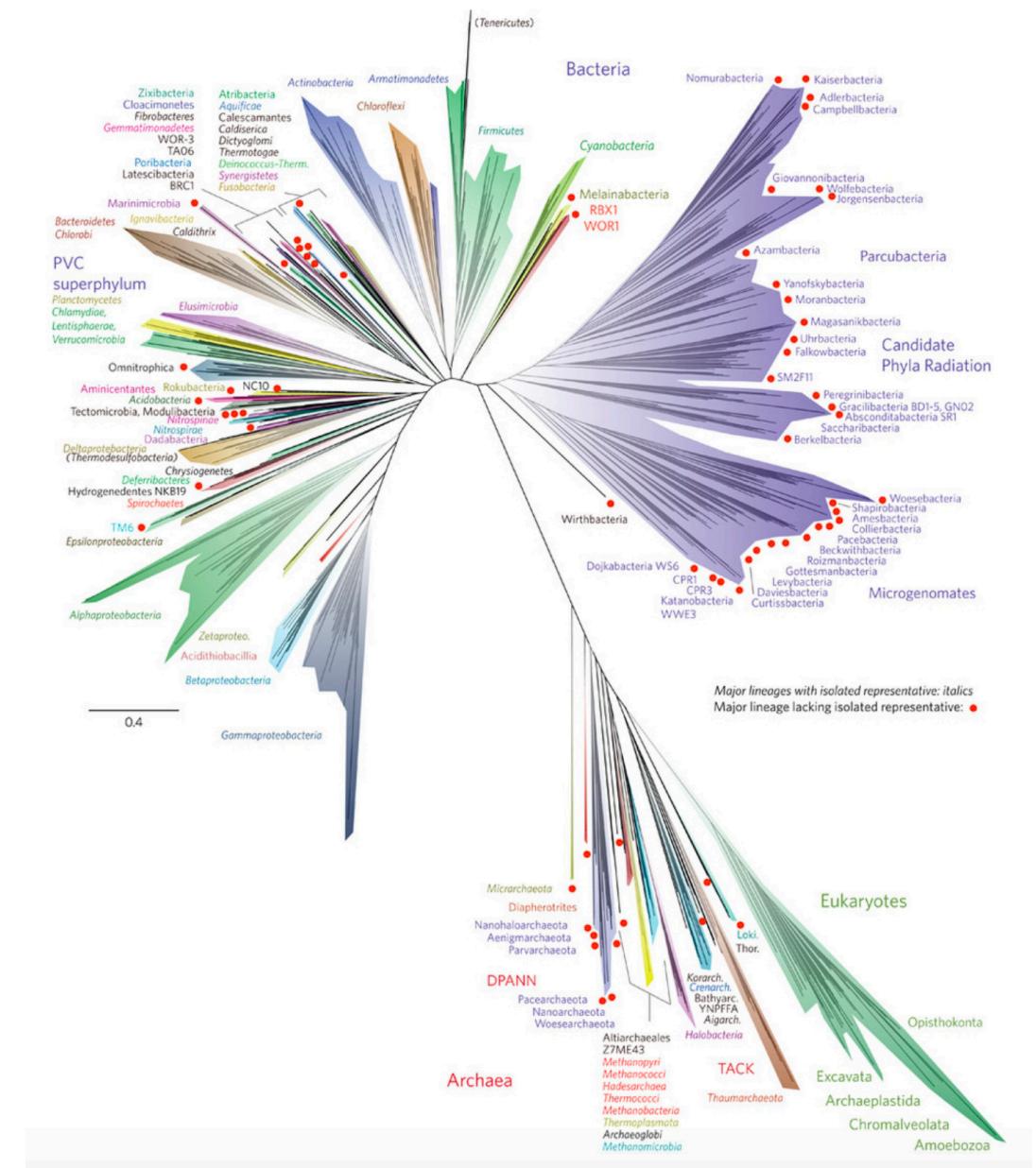
Molecular phylogenetics

- What are the derived traits?



Molecular phylogenetics

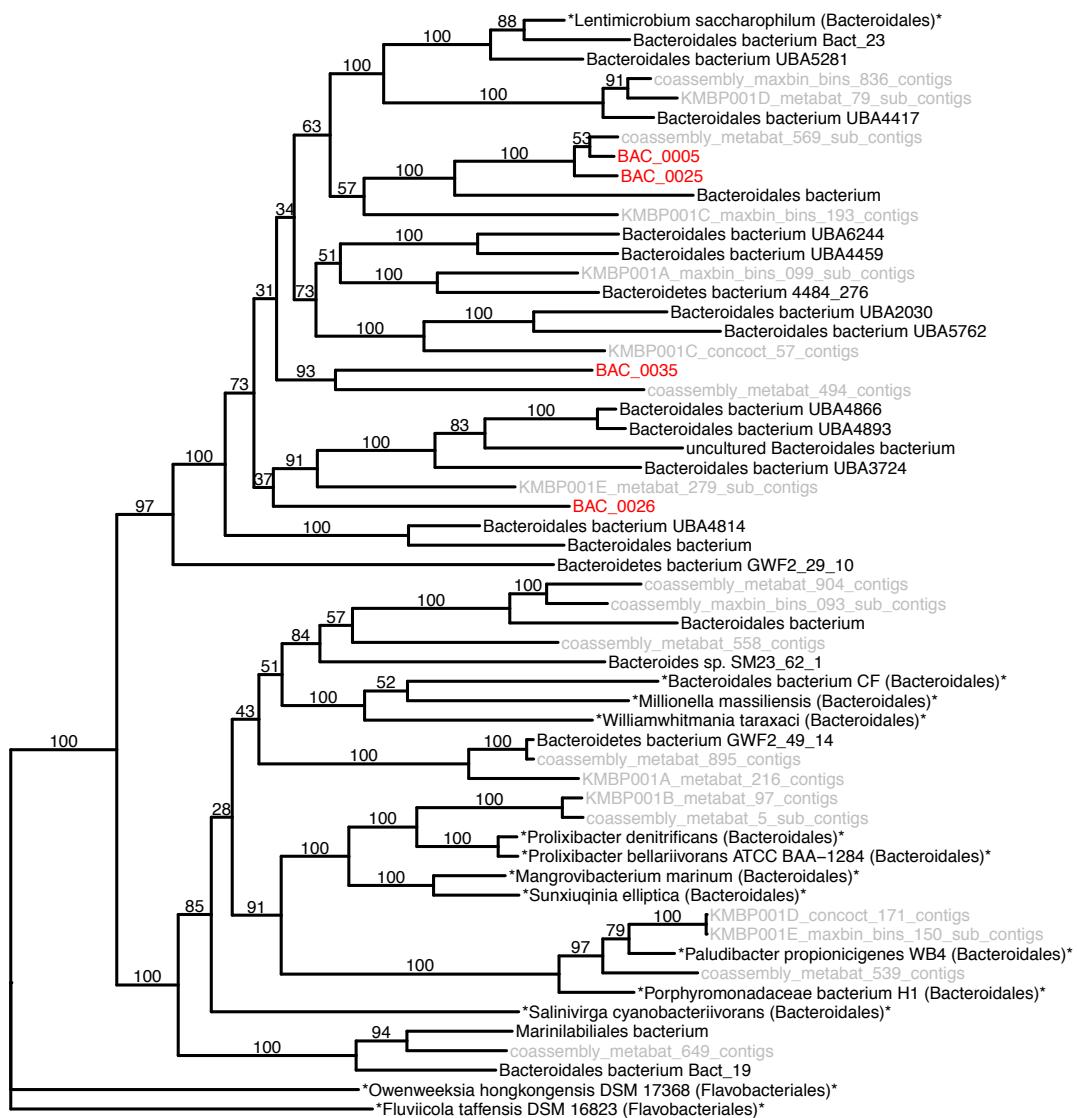
- Substitutions in sequences (nucleic acid or amino acids) are the “derived traits”
- Provide direct insight into actual evolutionary processes



Uses for phylogenetic inference

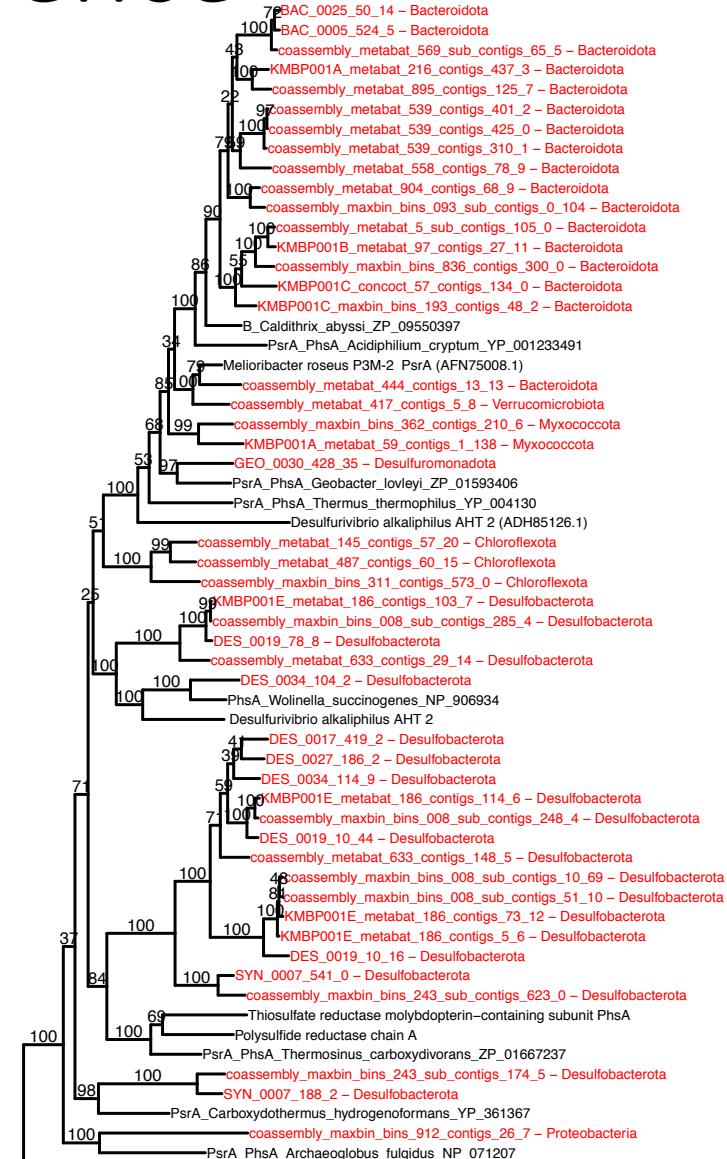
Uses for phylogenetic inference

- Taxonomic classification
- Evolutionary relationships of organisms/genes
 - Searching for horizontal gene transfer
- Estimate of functional classification of genes
- ??????????



Uses for phylogenetic inference

- Taxonomic classification
- Evolutionary relationships of organisms/genes
 - Searching for horizontal gene transfer
- Estimate of functional classification of genes
- ??????????



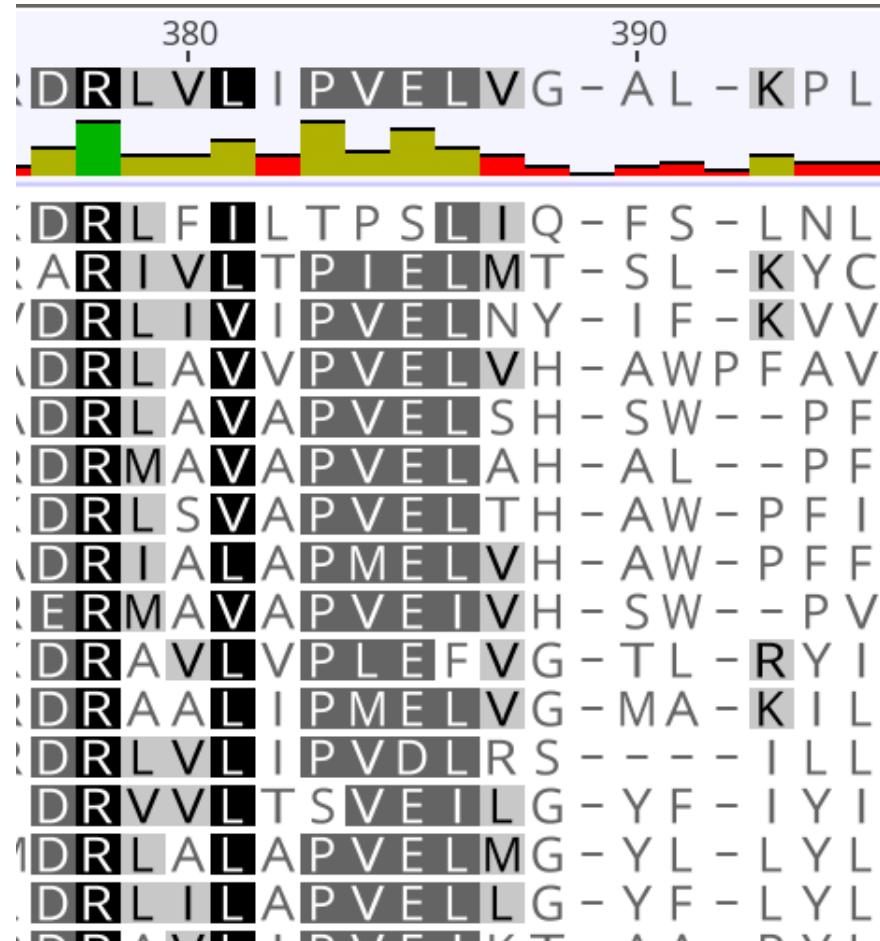
Steps in building a phylogenetic tree

1. Basic workflow
 1. Selection of sequence to use
 2. Generate sequence alignment
 1. Alignment masking and trimming
 3. Generate phylogenetic tree
 4. Check branch support and tree quality
 5. Root tree (if needed)
2. Potential modifications
 1. Removal of divergent sequences

Generation of sequence alignment

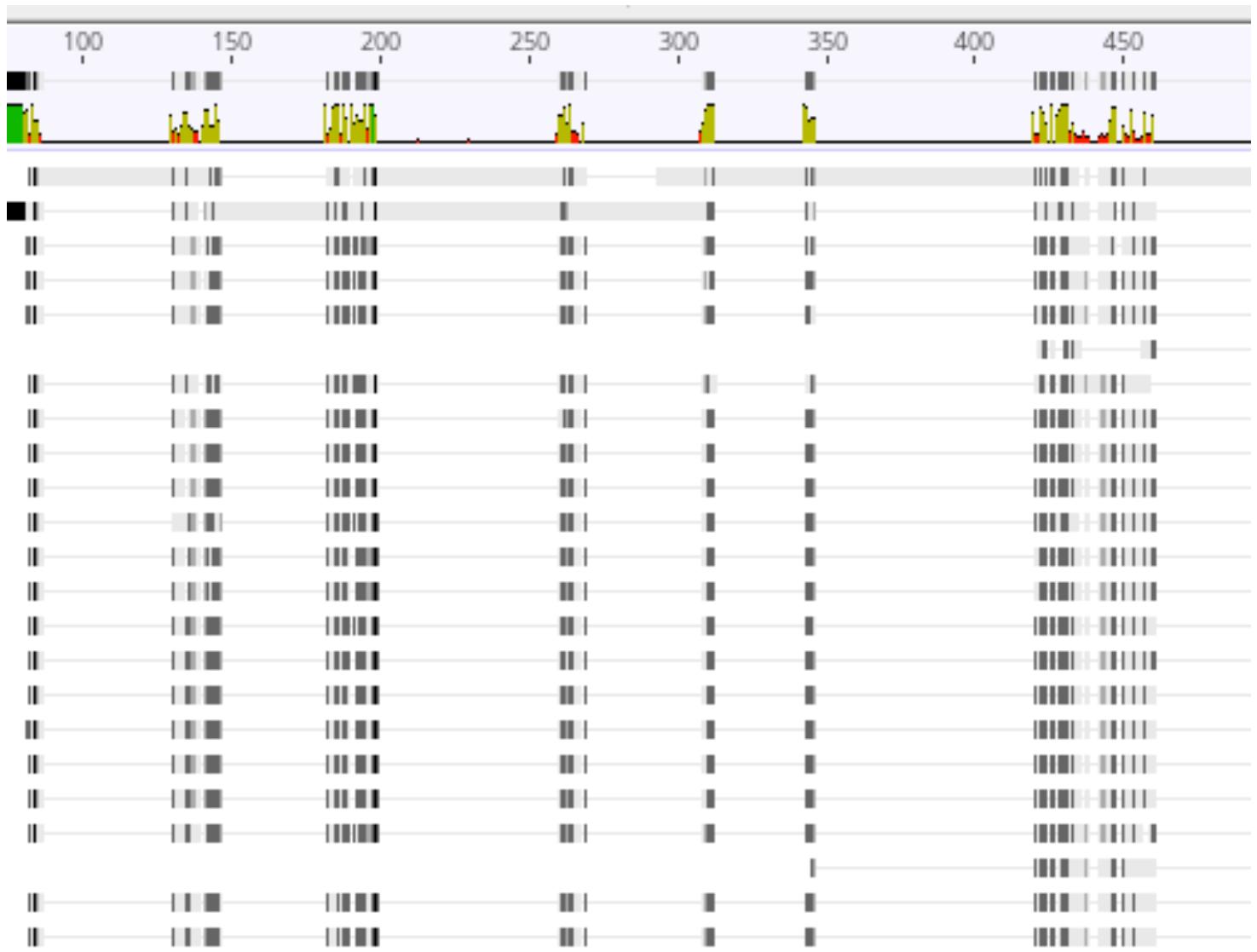
1. Multiple sequence alignment

1. MUSCLE
2. Clustal
3. T-Coffee
4. MAFFT



Generation of sequence alignment

1. Manual inspection of alignment
 1. Do some sequences need to be removed?

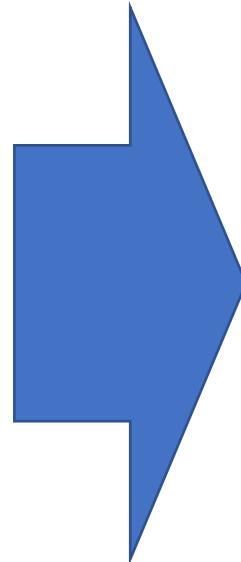


Masking Methods

1. Trimming strategies
 1. Fraction of gaps and/or trim overhanging ends
 2. Modeling suitability using substitution matrices (BMGE1.1)

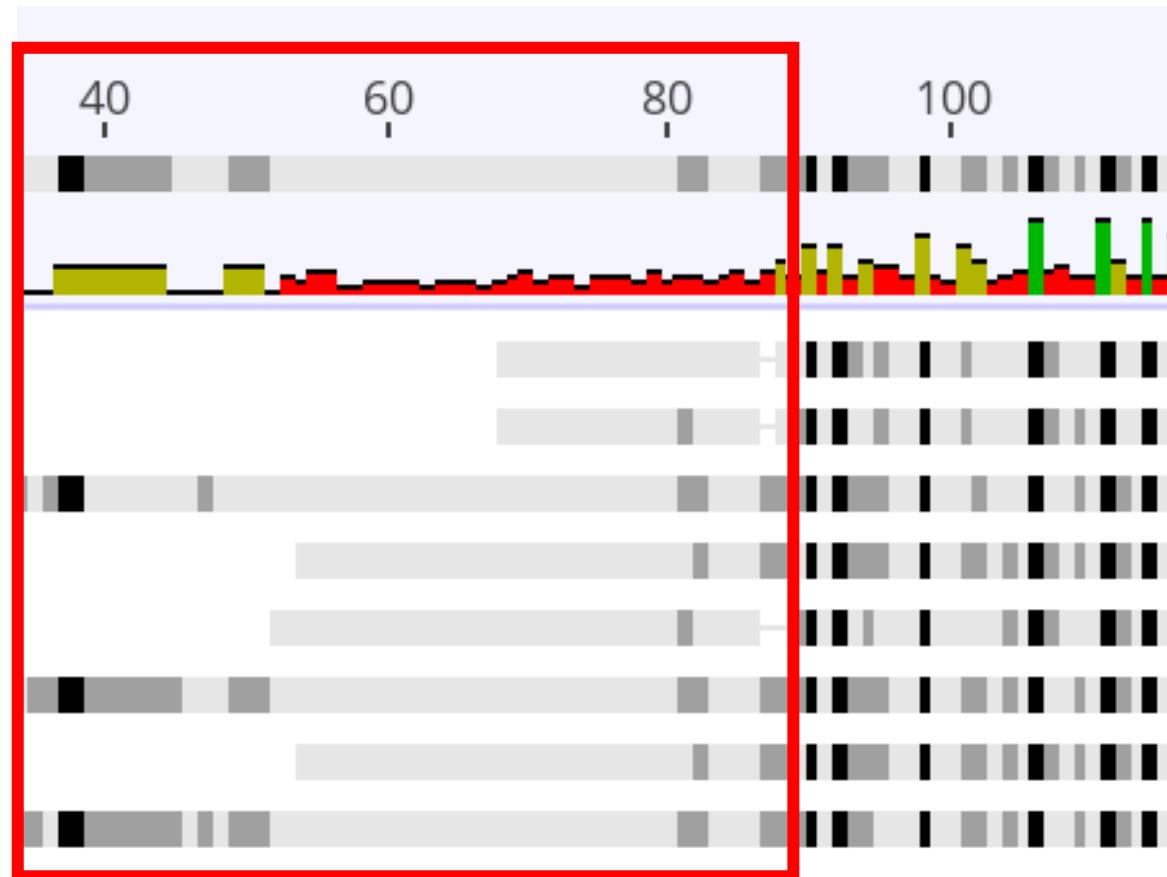
Masking an alignment

	380	390	
DRLVLI	PVELVG	G - A L - K P L	
DRLFIL	TPSLI	Q - F S - L N L	
ARIVL	TPIELMT	- S L - K Y C	
DRLIVI	PVELNY	- I F - K V V	
DRLAVV	PVELVH	- AWPFAV	
DRLAVV	PVELSH	- SW - P F	
DRMAVA	PVELAH	- AL - P F	
DRLSVV	PVELTH	- AW - P F I	
DRIALA	APMELV	H - AW - P F F	
ERMAVA	VAPVEIV	H - SW - P V	
DRAVVL	VPLEFV	G - TL - R Y I	
DRAALI	IPMELV	G - MA - K I L	
DRLVLI	IPVDLRS	- - - - I L L	
DRVVL	TSVEIL	G - Y F - I Y I	
DRLALA	APVELM	G - Y L - L Y L	
DRLIL	LAPVELL	G - Y F - L Y L	



	160	170	
DRLVLI	PVELVK	P L	
DRLFIL	TPSLIL	N L	
ARIVL	TPIELMK	Y C	
DRLIVI	PVELNK	V V	
DRLAVV	PVELV	F A V	
DRLAVV	PVELS	- P F	
DRMAVA	PVELA	- P F	
DRLSVV	PVELTP	F I	
DRIALA	APMELV	P F F	
ERMAVA	VAPVEIV	- P V	
DRAVVL	VPLEFV	R Y I	
DRAALI	IPMELV	K I L	
DRLVLI	IPVDLR	I L L	
DRVVL	TSVEIL	I Y I	
DRLALA	APVELM	L Y L	
DRLIL	LAPVELL	L Y L	

Trimming ends of alignment



Masking Methods

BMGE1.1 Report

1. Trimming strategies
 1. Fraction of gaps
 2. Overhanging ends
 3. Modeling suitability using substitution matrices
(BMGE1.1)

The diagram illustrates a sequence alignment with several features:

- Residue Count:** Below the sequence, four vertical bars indicate positions 190, 200, 210, and 220.
- Masking:** Regions of the sequence are masked with black squares. A large masked area starts at position 190 and extends to approximately position 225. Another smaller masked area is located between positions 210 and 220.
- Sequence Data:** The sequence consists of 15 lines of amino acid residues. The first few lines are:
 - DGLNAWLMVIDTRGVNVWCAAGKGNFSTNEVTEKLKKFKVS
 - DGMDAWILVLDTKGINVWCAAGKGTFGTGEIISKIKEFKLG
 - GGTNLWILVVDTKGIVNVWCAAGKGNFGTDVVVRGIKKTSL
 - DGIDAWILVLDTKGVNVWCAAGKGTFGTDELVYRIASVSLN
 - QGYSAWLLVLDTKGVNVWCAAGKGTFGTEEVIQKLESSQLH
 - KGYPVWILVLDTKGINVWCAAGKGTFGTEELIHQIESADIKI
 - NSISAWILVLDSEGVNVWCAAGKGTFSQELVGRISKSAVEI
 - KGQNLWILVLDTKGVNVWCAAGKGTFSCELVNRIHASELE
 - SSINAWILVLDTKGINVWCAAGKGTFSHELVKRIKKSSLE
 - PGINAWILVLNTNGINVWCAAGKGTFSRELANKIKDASLE
 - AGLDAWLLVVDTRGINVWCAAGKGTFCAEVARVVREVSLA
 - AGLDAWLLVIDSRGINVWCAAGKGTFSSEEIAYQVQRCLAI

Steps in building a phylogenetic tree

1. Basic workflow

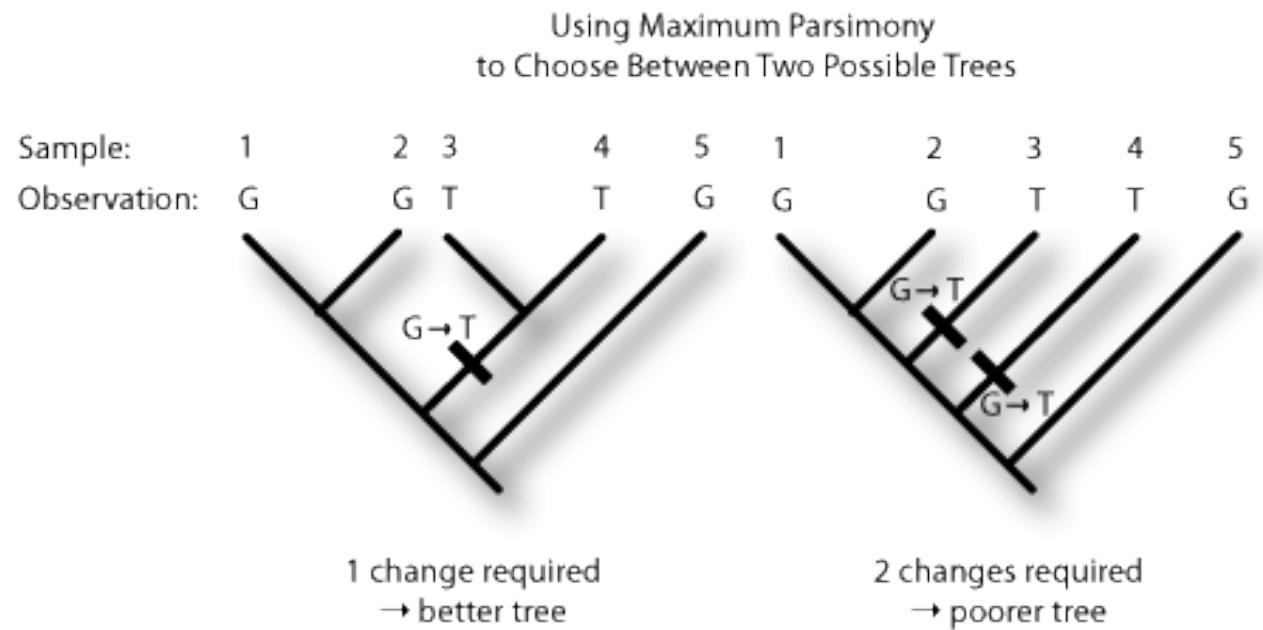
1. Selection of sequence to use
2. Generate sequence alignment
 1. Alignment masking and trimming
3. Generate phylogenetic tree
4. Check branch support and tree quality
5. Root tree (if needed)

2. Potential modifications

1. Removal of divergent sequences

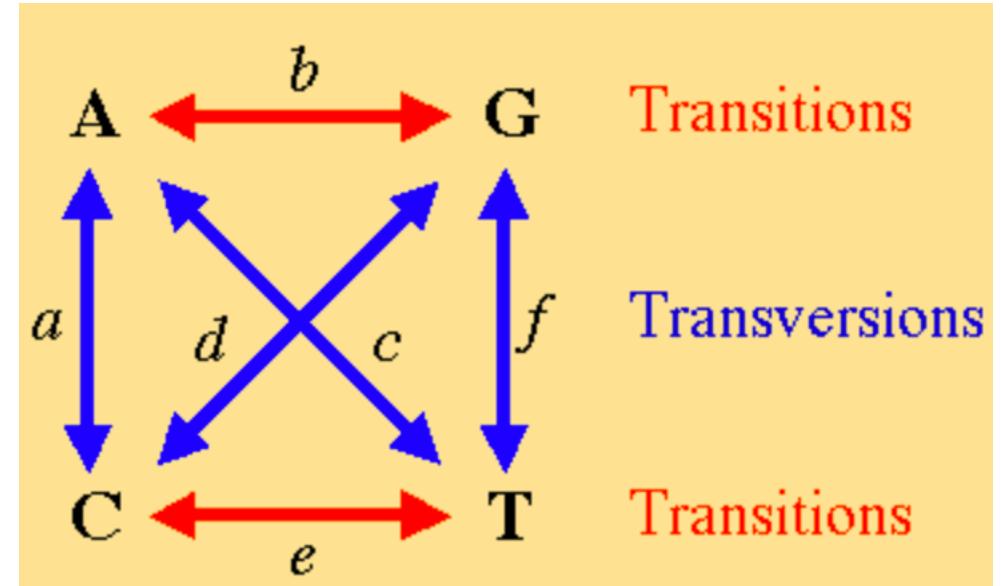
Common methods of tree generation

1. Maximum Parsimony
 1. No model for evolution
 2. Tries to minimize number of changes in sequence
 3. Not often used these days



Molecular models of evolution - DNA

1. Need to account for likelihood of shifts in sequences
 1. Jukes-Cantor assumes all are equally likely
 2. Others have differential rates



Molecular models of evolution - proteins

1. Protein substitutions are significantly more complicated
 1. Many different models: LG, WAG, BLOSUM, etc etc

Ala	4																		
Arg	-1 5																		
Asn	-2 0 6																		
Asp	-2 -2 1 6																		
Cys	0 -3 -3 -3 9																		
Gln	-1 1 0 0 -3 5																		
Glu	-1 0 0 2 -4 2 5																		
Gly	0 -2 0 -1 -3 -2 -2 6																		
His	-2 0 1 -1 -3 0 0 -2 8																		
Ile	-1 -3 -3 -3 -1 -3 -3 -4 -3 4																		
Leu	-1 -2 -3 -4 -1 -2 -3 -4 -3 2 4																		
Lys	-1 2 0 -1 -3 1 1 -2 -1 -3 -2 5																		
Met	-1 -1 -2 -3 -1 0 -2 -3 -2 1 2 -1 5																		
Phe	-2 -3 -3 -3 -2 -3 -3 -3 -1 0 0 -3 0 6																		
Pro	-1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4 7																		
Ser	1 -1 1 0 -1 0 0 0 -1 -2 -2 0 -1 -2 -1 4																		
Thr	0 -1 0 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1 1 5																		
Trp	-3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1 1 -4 -3 -2 11																		
Tyr	-2 -2 -2 -3 -2 -1 -2 -3 2 -1 -1 -2 -1 3 -3 -2 -2 2 7																		
Val	0 -3 -3 -3 -1 -2 -2 -3 -3 3 1 -2 1 -1 -2 -2 0 -3 -1 4																		
Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

<https://upload.wikimedia.org/wikipedia/commons/0/02/BLOSUM62.png>

Common methods of tree generation

1. Distance methods
 1. Quick, useful for checking tree
 2. Generates distance matrix from alignment
 3. **Neighbor-joining method**
2. Maximum Likelihood
 1. Computationally intensive, but shortcuts can be made
 2. Generates multiple optimized trees and compares to model-predicted score
 3. **RAXML or FastTree (approximately)**
3. Bayesian inference
 1. Computationally intensive, complex, requires accurate prior probabilities
 2. Returns actual probability of data being correct (based on priors)
 3. Allows for incorporation of prior knowledge
 4. **Mr. Bayes**

Steps in building a phylogenetic tree

1. Basic workflow

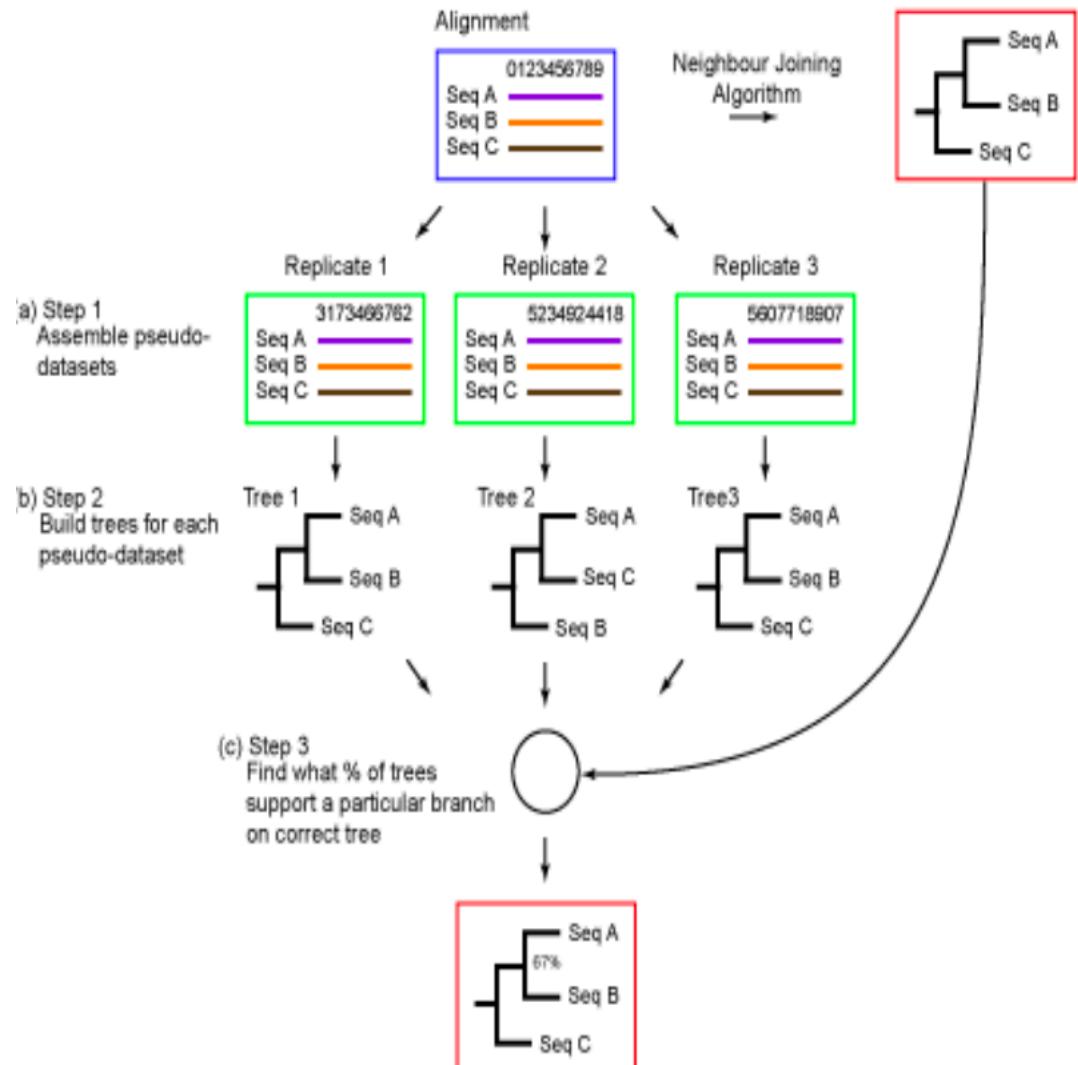
1. Selection of sequence to use
2. Generate sequence alignment
 1. Alignment masking and trimming
3. Generate phylogenetic tree
4. Check branch support and tree quality
5. Root tree (if needed)

2. Potential modifications

1. Removal of divergent sequences

Branch support methods - Bootstrapping

1. Random subsampling of the alignment with replacement
2. Generation of tree based on new sequences
3. Repeat many times
4. Calculate fraction of trees that have a particular branch
5. **Is a branch label, not a node label**



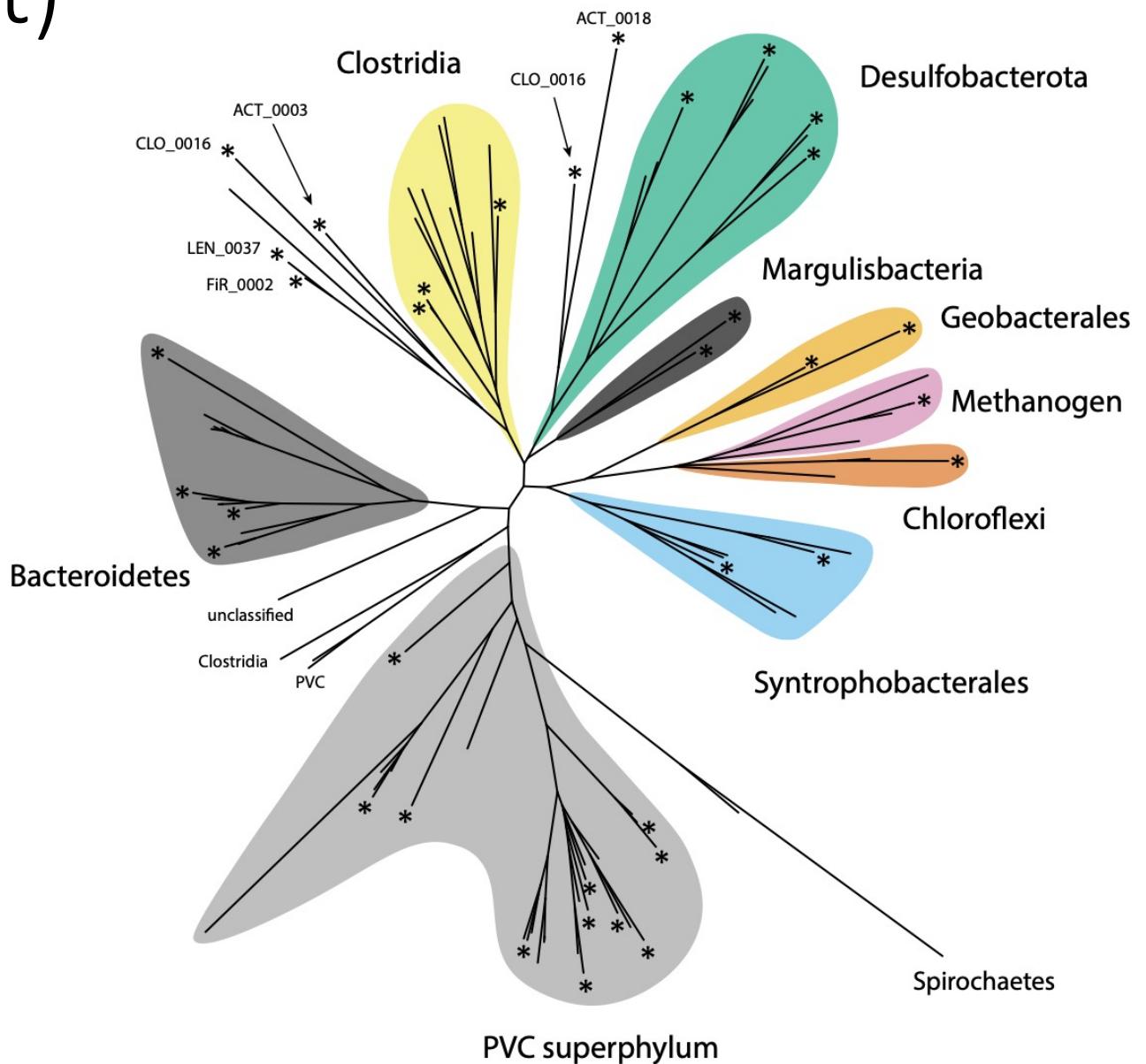
To root, or not to root?

1. Main points to consider:
 1. What is your question?
 1. Relatedness vs. evolution
 2. What information do you have about the sequences and their evolutionary history?



Tree rooting (or not)

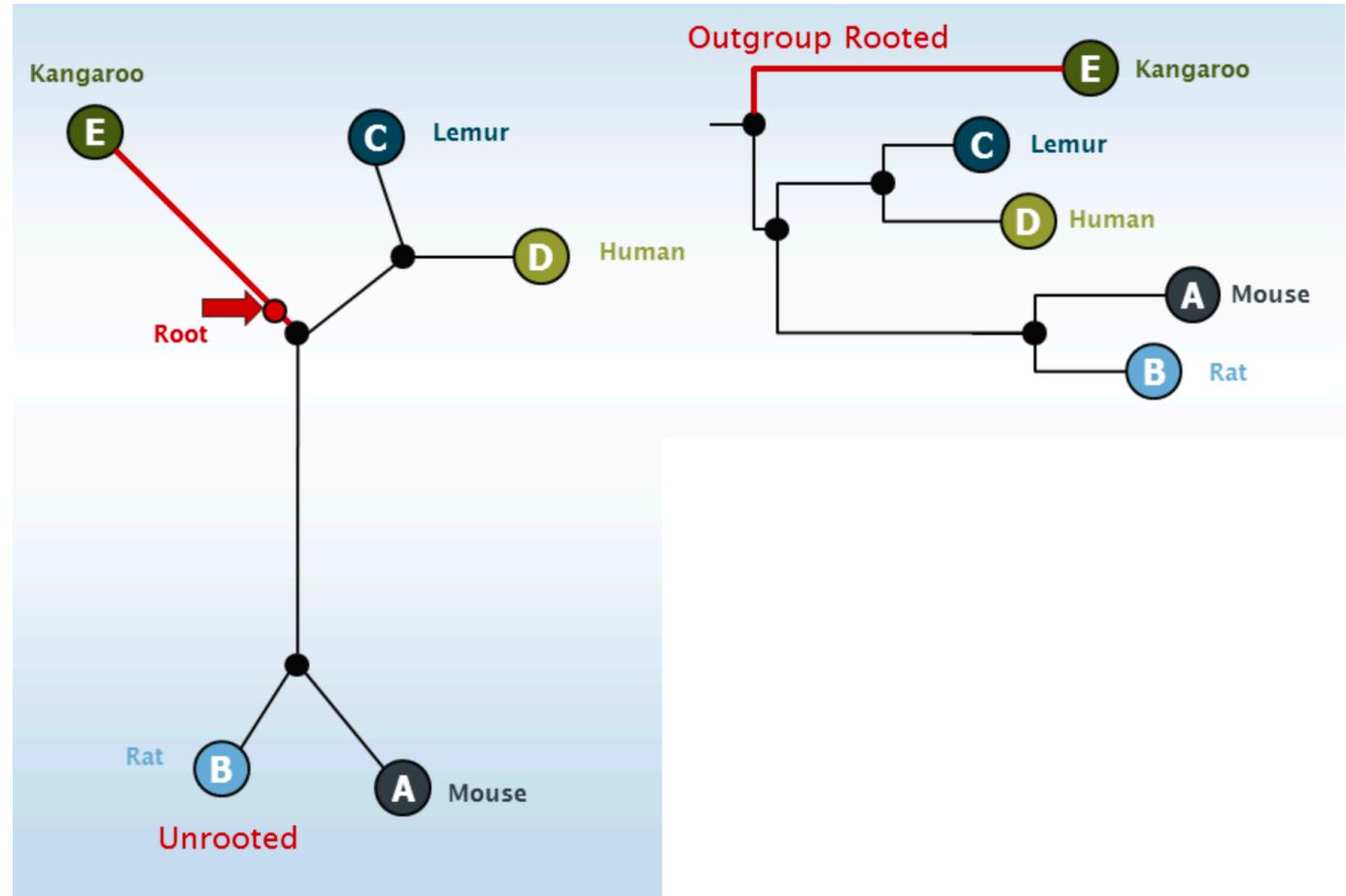
1. Leave it unrooted
 1. Useful if just interested in the “similarity” or “relatedness”
 2. Can’t be used to discuss ancestral state or direction of evolution



Tree rooting (or not)

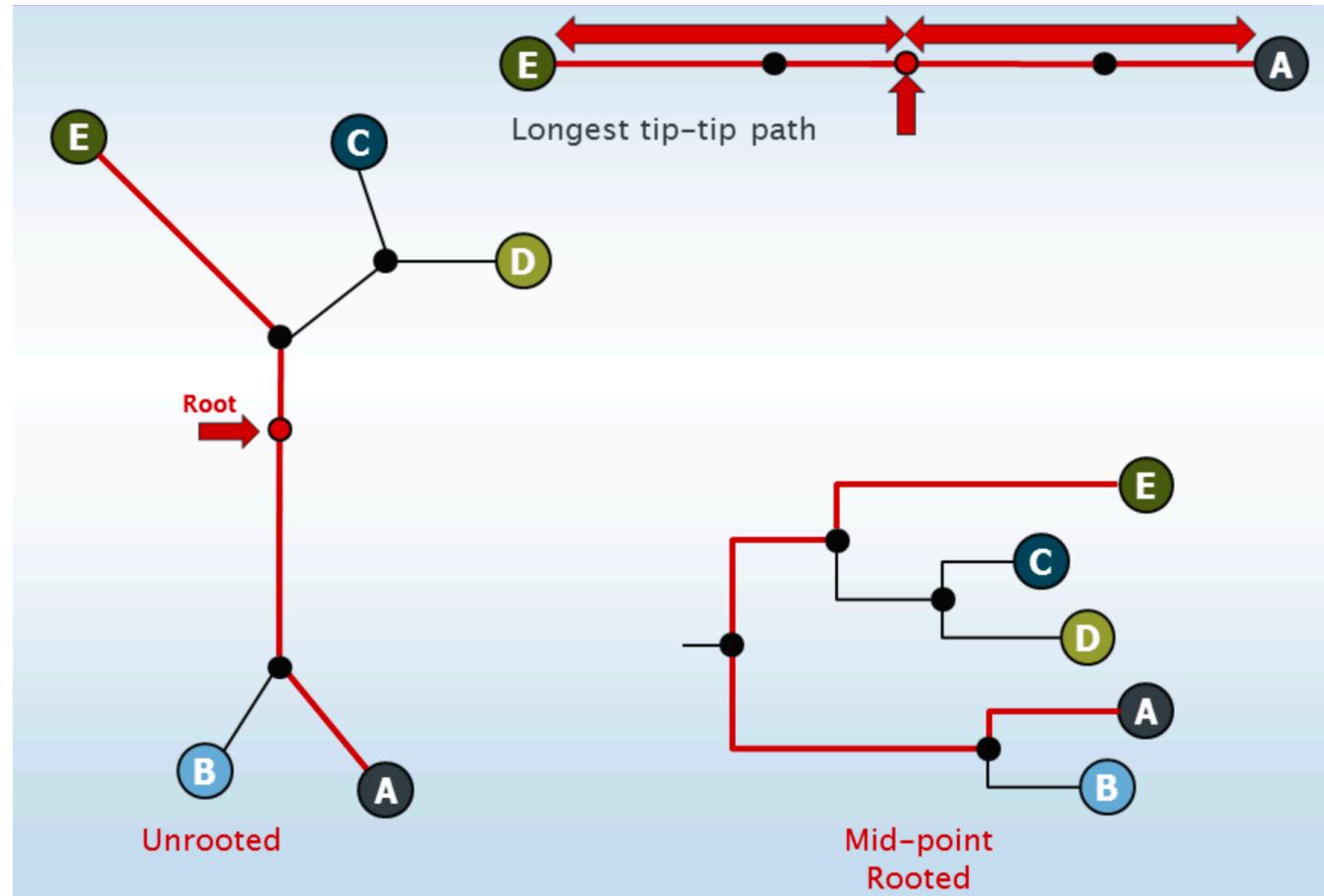
2. Outgroup rooting

1. Assumes you know a true outgroup and that it's closely related enough
2. Can include outgroups in the alignment and tree generation
3. Or can generate tree without outgroup and place the outgroup on the tree later



Tree rooting (or not)

3. Midpoint rooting
 1. Assumes constant evolutionary rate
 2. Doesn't require *a priori* knowledge of outgroup



Reiteration of tree

1. Done if you want to improve bootstrapping values/posterior probabilities
2. Removal of highly divergent sequences
 1. Manual inspection or automatic (RogueNaRok)
3. Different masking techniques

References to check out

1. **Tree Thinking – David Baum and Stacey Smith**
2. Papers by Alexi Stamatakis
3. Molecular phylogenetics: principles and practice (review paper)
4. <http://phylobotanist.blogspot.com/2015/01/how-to-root-phylogenetic-tree-outgroup.html>
5. <http://cabbagesofdoom.blogspot.com/2012/06/how-to-root-phylogenetic-tree.html>

I think



Darwin's Tree